機械学習における最適化理論と学習理論的側面

第二部: 非凸確率的最適化と再生核ヒルベルト空間の最適化

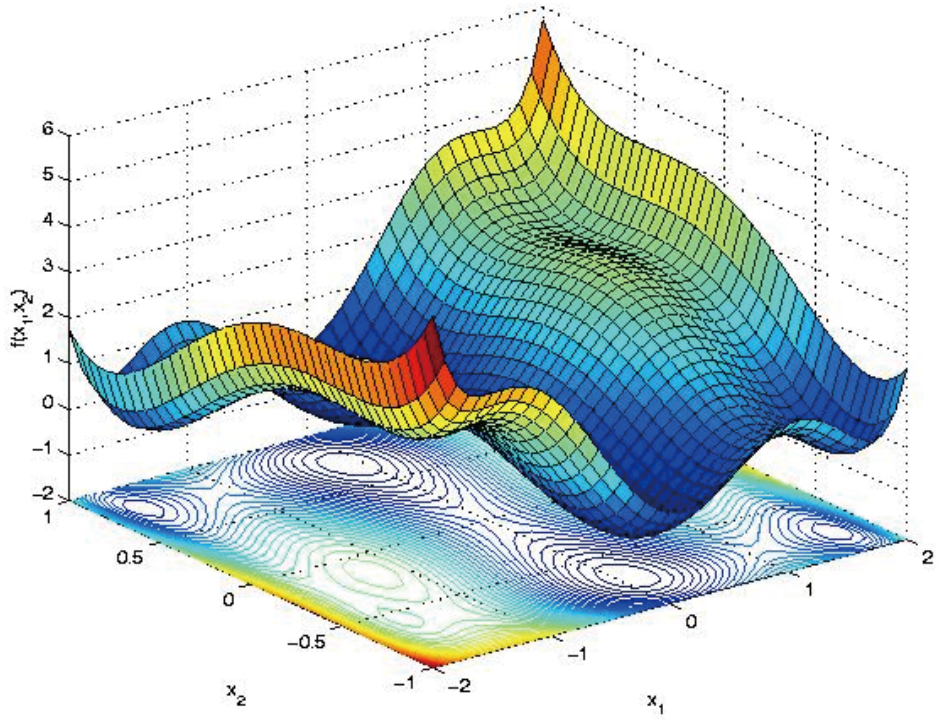
鈴木大慈

東京大学大学院情報理工学系研究科数理情報学専攻
理研 AIP

2020 年 8 月 6 日
@組合せ最適化セミナー 2020 (COSS2020)

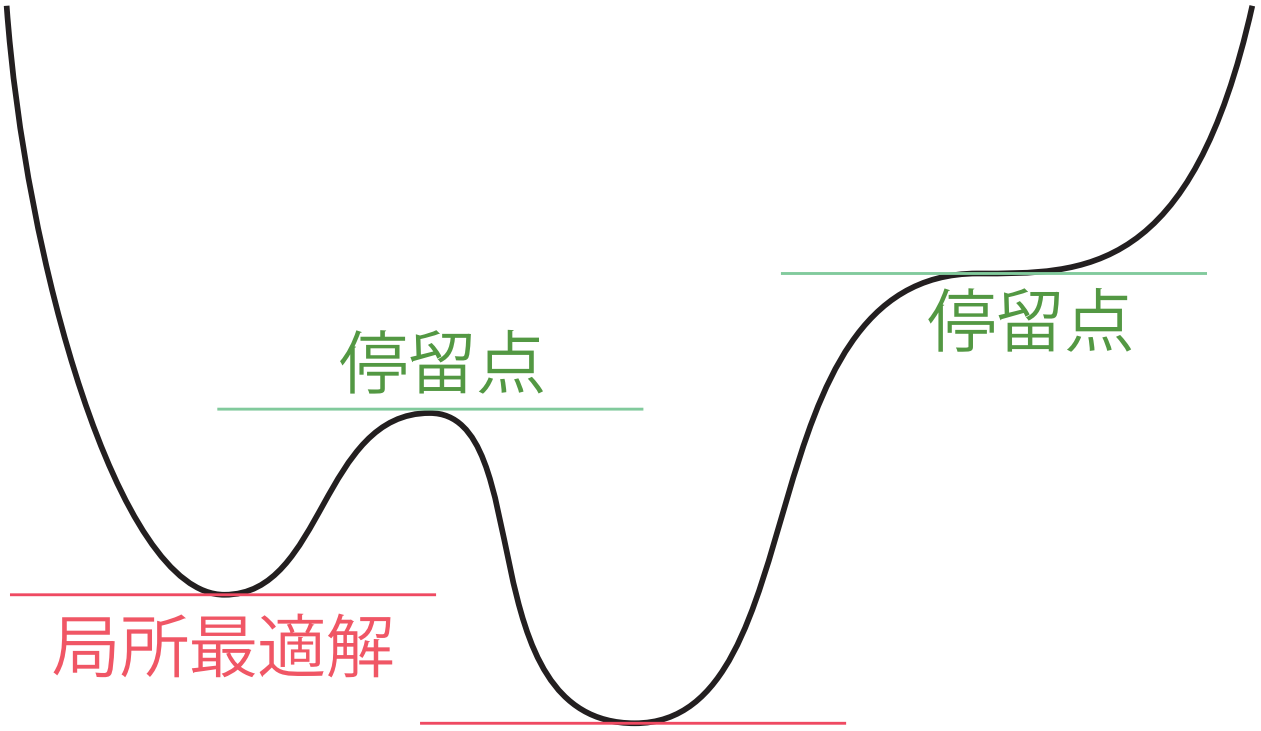
Outline

非凸関数最小化



- これまで紹介した凸最適化手法をそのまま当てはめても実用上は結構動く．
- ただし，双対問題を解く方法はそのままでは適用できない．
- ステップサイズを適切に選ぶ必要がある．
- 大域的最適化は難しい．微分が 0 になる停留点への収束は保証できる．
- 大域的最適化を厳密に行うにはアニーリングなどの方法を使う必要がある．

停留点と局所最適解



→大域的最適解でもある

非凸関数での SGD

目的関数：$L(x) = \mathbb{E}_z[\ell(z, x)]$.　(下に有界, $L^* = \inf_x L(x)$ とする)

SGD

- $z_t \sim P(Z)$ を観測．$\ell_t(x) := \ell(z_t, x)$ とする．
- $g_t \in \partial_x \ell_t(x_{t-1})$.
- $x_t = x_{t-1} - \eta_t g_t$.

仮定

(A1) L は γ-平滑

(A2) $\mathbb{E}[\|g_t - \mathbb{E}[g_t]\|^2] = \sigma^2$ (確率的勾配の分散は σ^2).

- $\eta_t = \min\left\{\frac{\tilde{D}}{\sigma\sqrt{T}}, \frac{1}{\gamma}\right\}$ とすると (Ghadimi and Lan (2013))

$$\min_{1\le t\le T} \mathbb{E}[\|\nabla L(x_t)\|^2] \le \frac{\gamma\sigma}{\sqrt{T}}\left(\frac{D_f^2}{\tilde{D}} + \tilde{D}\right) + \frac{\gamma^2 D_f^2}{T},$$

ただし，$D_f = \sqrt{\frac{2(L(x_1)-L^*)}{\gamma}}$.　（微分が 0 へ収束してゆくことを保証）

左辺の $\min_{1\le t\le T}$ の代わりに，$\hat{t} \in \{1, \dots, T\}$ を一様分布に従って選んで $\mathbb{E}[\|\nabla L(x_{\hat{t}})\|^2]$ としても良い．

非凸 SVRG

$$\min_{x\in\mathbb{R}^p} L(x) = \min_{x\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \ell_i(x)$$

SVRG をそのまま非凸関数最適化に適用してよい.（ただしステップサイズとミニバッチ数は適切に調整）

$\mathbb{E}[\|\nabla L(\hat{x})\|^2] \leq \epsilon$ になるまでの更新回数 T (Allen-Zhu and Hazan, 2016, Reddi et al., 2016)

- ℓ_i がγ-平滑の時：

$$T \geq \Omega\left(n + \frac{n^{2/3}}{\epsilon}\right).$$

（普通の非確率的勾配法なら $\Omega(n/\epsilon)$）

- ℓ_i が γ-平滑かつ $L(x) - L(x^*) \leq \tau\|\nabla L(x)\|^2$ $(\forall x)$ (x^* は大域的最適解) の時 (Polyak- Lojasiewicz, PL 条件)：

$$T \geq \Omega\left((n + \tau n^{2/3})\log(1/\epsilon)\right).$$

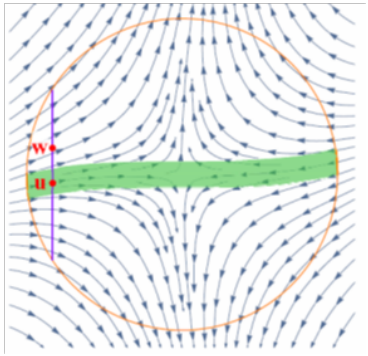
（普通の非確率的勾配法なら $\Omega(\tau n \log(1/\epsilon))$）

鞍点回避の方法

- 単純な勾配法に雑音を乗せる (Jin et al., 2017a)

for $t = 0, 1, \ldots$ **do**
  if perturbation condition holds **then**
    $\mathbf{x}_t \leftarrow \mathbf{x}_t + \xi_t, \quad \xi_t$ uniformly $\sim \mathbb{B}_0(r)$
  $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t)$ (普通の勾配法)

ノイズを乗せるだけ
(鞍点脱出)



- 加速勾配法への適用 (Jin et al., 2017b)

1: $\mathbf{v}_0 \leftarrow 0$
2: **for** $t = 0, 1, \ldots,$ **do**
3: **if** $\|\nabla f(\mathbf{x}_t)\| \le \epsilon$ and *no perturbation in last* $\mathcal{T}$ *steps* **then** } Perturbation
4: $\mathbf{x}_t \leftarrow \mathbf{x}_t + \xi_t \quad \xi_t \sim \mathrm{Unif}(\mathbb{B}_0(r))$
5: $\mathbf{y}_t \leftarrow \mathbf{x}_t + (1-\theta)\mathbf{v}_t$ } 加速勾配法 } AGD
6: $\mathbf{x}_{t+1} \leftarrow \mathbf{y}_t - \eta \nabla f(\mathbf{y}_t)$
7: $\mathbf{v}_{t+1} \leftarrow \mathbf{x}_{t+1} - \mathbf{x}_t$
8: **if** $f(\mathbf{x}_t) \le f(\mathbf{y}_t) + \langle \nabla f(\mathbf{y}_t), \mathbf{x}_t - \mathbf{y}_t \rangle - \frac{\gamma}{2}\|\mathbf{x}_t - \mathbf{y}_t\|^2$ **then** } Negative curvature
9: $(\mathbf{x}_{t+1}, \mathbf{v}_{t+1}) \leftarrow$ Negative-Curvature-Exploitation$(\mathbf{x}_t, \mathbf{v}_t, s)$ } exploitation

鞍点にはまっていそうならノイズを乗せて脱出

凸っぽくなければある方法で降下方向を発見

SARAH とその改良法

SSRGD (Li, 2019): SARAH + ノイズ付加による鞍点脱出

Simple Stochastic Recursive Gradient Descent (SSRGD)

Iterate the following for $t = 1, 2, \ldots, T$:

1. 鞍点脱出モードに入っておらず，$\|\nabla L(x_t)\| \leq g_{\mathrm{thresh}}$ なら，
   - $x_t \leftarrow x_t + \xi$ $(\xi \sim \mathrm{Unif}(B_r(\mathbb{R}^d)))$ として，鞍点脱出モードに入る．
2. $y_0 = x_t$, $v_0 = \nabla f(x_t)$
3. For $k = 1, \ldots, m$,
   1. $y_k = y_{k-1} - \eta v_{k-1}$
   2. $v_k = \frac{1}{b} \sum_{i \in I_k} (\nabla f_i(y_k) - \nabla f_i(y_{k-1})) + v_{k-1}$ (**SARAH**: variance reduction)
   3. ある停止条件を満たしていたら鞍点脱出モードを止める．
4. $x_{t+1} = y_m$

Output: x_T

- SARAH: StochAstic Recursive grAdient algoritHm (Nguyen et al., 2017, Pham et al., 2020)
- オンライン型の場合は ∇L の計算はサンプル平均にする $\frac{1}{B} \sum_{i \in I_t} \nabla f_i(x_t)$.
- 二次最適性も高い確率で保証

SARAH について

- SARAH: $v_k = \frac{1}{b}\sum_{i\in I_k}(\nabla f_i(y_k) - \nabla f_i(y_{k-1})) + v_{k-1}$
- SVRG: $v_k = \frac{1}{b}\sum_{i\in I_k}(\nabla f_i(y_k) - \nabla f_i(\tilde{x})) + \tilde{v}$

SVRG は内部ループの更新を進めると分散が大きくなる.
SARAH は内部ループの更新を進めても分散が大きくならない or 0 に収束する
(強凸の場合)
→ 勾配が暴れず，一時最適性条件を満たす解を得やすい.
(凸最適化で目的関数値を見ている限りはこの違いが見にくい)

計算量の比較

- ϵ-一次最適性条件: $\mathbb{E}[\|\nabla L(x)\|^2] \leq \epsilon$
- δ-二次最適性条件: $\lambda_{\min}(\nabla^2 L(x)) \geq -\delta$ (with high probability)

オンライン型

手法	確率的勾配の計算数	最適性条件
GD	$O(\frac{n}{\epsilon})$	1 次
SVRG (Allen-Zhu and Hazan, 2016)	$O(n + \frac{n^{2/3}}{\epsilon})$	1 次
SARAH (Pham et al., 2020)	$O(n + \frac{\sqrt{n}}{\epsilon})$	1 次
SSRGD (Li, 2019)	$O(n + \frac{\sqrt{n}}{\epsilon})$	1 次
PGD (Jin et al., 2017b)	$O(\frac{n}{\epsilon} + \frac{n}{\delta^4})$	2 次
SSRGD (Li, 2019)	$O(\frac{\sqrt{n}}{\epsilon} + \frac{\sqrt{n}}{\delta^4} + \frac{n}{\delta^3})$	2 次

有限和型

手法	確率的勾配の計算数	最適性条件
SGD (Ghadimi and Lan, 2013)	$O(1/\epsilon^2)$	1 次
SVRG+ (Li and Li, 2018)	$O(1/\epsilon^{7/4})$	1 次
SARAH (Pham et al., 2020)	$O(1/\epsilon^{3/2})$	1 次
SSRGD (Li, 2019)	$O(1/\epsilon^{3/2})$	1 次
SSRGD (Li, 2019)	$O(\frac{1}{\epsilon^{3/2}} + \frac{1}{\epsilon\delta^3} + \frac{1}{\epsilon^{1/2}\delta^4})$	2 次

（参考）Strict saddle

- 深層学習などは停留点が多い．
- 目的関数が strict saddle property という性質を満たしていれば，サドルポイントを回避することができる．

信頼領域法 (Conn et al., 2000) や雑音を加えた確率的勾配法 (Ge et al., 2015) は strict saddle な目的関数の局所的最適解に到達する (Sun et al., 2015).

※ 解に雑音を加えることでサドルポイントから抜け出せる．

Strict saddle

二回微分可能な関数 f が *strict saddle* であるとは，$\forall x$ で次のどれかが満たされている:

- $\|\nabla f(x)\| \geq \epsilon$.
- $\lambda_{\min}(\nabla^2 f(x)) \leq -\gamma$.
- ある x^* が存在して $\|x - x^*\| \leq \delta$ かつ $f(x)$ が x^* の近傍 $\{x' \mid \|x^* - x'\| \leq 2\delta\}$ で強凸関数.

E.g., テンソル分解 $\max_{u \in \mathbb{R}^p} \left\langle \sum_{r=1}^d a_r^{\otimes 4}, u \otimes u \otimes u \otimes u \right\rangle$ は $a_r^\top a_{r'} = \delta_{r,r'}$ なら strict saddle.

線形制約ありの学習問題

$$\min_x \frac{1}{n}\sum_{i=1}^n f_i(z_i^\top x) + \psi(B^\top x)$$

$$\Leftrightarrow \quad \min_{x,y} \frac{1}{n}\sum_{i=1}^n f_i(z_i^\top x) + \psi(y) \ \text{ s.t. } \ y = B^\top x.$$

拡張ラグランジアン

$$\mathcal{L}(x, y, \lambda) = \frac{1}{n}\sum_i f_i(z_i^\top x) + \psi(y) + \lambda^\top (y - B^\top x) + \frac{\rho}{2}\|y - B^\top x\|^2$$

$$\inf_{x,y} \sup_{\lambda} \mathcal{L}(x, y, \lambda)$$

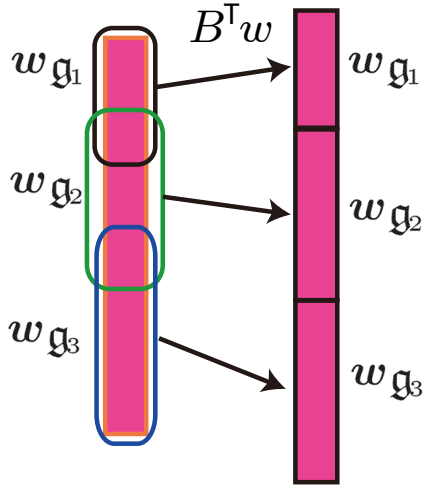
で最適解が求まる.

- 乗数法: Hestenes (1969), Powell (1969), Rockafellar (1976).
- 交互方向乗数法 (ADMM): Gabay and Mercier (1976), Mota et al. (2011), He and Yuan (2012), Deng and Yin (2012), Hong and Luo (2012a)
- 確率的交互方向乗数法: SGD-ADMM (Suzuki, 2013, Ouyang et al., 2013), RDA-ADMM (Suzuki, 2013), SDCA-ADMM (Suzuki, 2014), SVRG-ADMM (Zheng and Kwok, 2016), ASVRG-ADMM (Liu et al., 2017).

構造的正則化の例

- Overlapped group lasso $\tilde{\psi}(w) = C \sum_{\mathfrak{g} \in \mathfrak{G}} \|w_{\mathfrak{g}}\|$

It is difficult to compute the proximal mapping.



Solution:

- Prepare ψ for which proximal mapping is easily computable.
- Let $\psi(B^\top w) = \tilde{\psi}(w)$, and utilize the proximal mapping w.r.t. ψ.

Decompose into independent groups:

$$B^\top w = \begin{bmatrix} w_{\mathfrak{g}_1} \\ w_{\mathfrak{g}_2} \\ w_{\mathfrak{g}_3} \end{bmatrix}$$

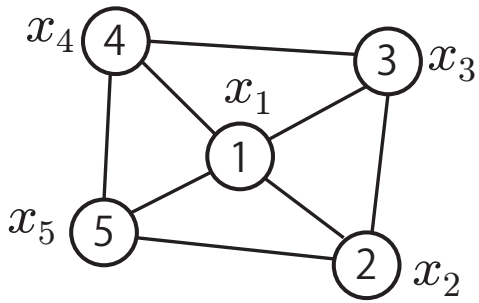
$$\psi(y) = C \sum_{\mathfrak{g}' \in \mathfrak{G}'} \|y_{\mathfrak{g}'}\|$$

$$\mathrm{prox}(q|\psi) = \left(q_{\mathfrak{g}'} \max\left\{ 1 - \frac{C}{\|q_{\mathfrak{g}'}\|}, 0 \right\} \right)_{\mathfrak{g}' \in \mathfrak{G}'}$$

その他の例

- Graph guided regularization

$$\tilde{\psi}(w) = C \sum_{(i,j)\in E} |w_i - w_j|.$$



$$\psi(y) = C \sum_{e\in E} |y_e|, \quad y = B^\top w = (w_i - w_j)_{(i,j)\in E}$$

$$\Rightarrow \begin{cases} \psi(B^\top w) = \tilde{\psi}(w), \\ \mathrm{prox}(q|\psi) = \left(q_e \max\left\{1 - \frac{C}{|q_e|}, 0\right\}\right)_{e\in E}. \end{cases}$$

Soft-Thresholding function.

構造的正則化に対する交互方向乗数法

$$\min_x \{f(x) + \psi(B^\top w)\} \Leftrightarrow \min_{x,y} \{f(x) + \psi(y) \text{ s.t. } y = B^\top x\}$$

$$\mathcal{L}(x, y, \lambda) = f(x) + \psi(y) + \lambda^\top (y - B^\top x) + \tfrac{\rho}{2}\|y - B^\top x\|^2$$

ただし $f(x) = \frac{1}{n}\sum f_i(z_i^\top x)$

ADMM による構造的正則化学習

$$x^{(t)} = \arg\min_x \{f(x) + {\lambda^{(t-1)}}^\top (-B^\top x) + \frac{\rho}{2}\|y^{(t-1)} - B^\top x\|^2\}$$

$$y^{(t)} = \arg\min_y \{\psi(y) + {\lambda^{(t)}}^\top y + \frac{\rho}{2}\|y - B^\top x^{(t)}\|^2\}$$

$$(= \mathrm{prox}(B^\top x^{(t)} - \lambda^{(t)}/\rho | \psi/\rho))$$

$$\lambda^{(t)} = \lambda^{(t-1)} - \rho(B^\top x^{(t)} - y^{(t)})$$

- y の更新は単純な ψ による近接写像.
  → 解析解.
- 一般的には $O(1/k)$ (He and Yuan, 2012), 強凸ならば線形収束 (Deng and Yin, 2012, Hong and Luo, 2012b).

SGD-ADMM

$\min_x \mathbb{E}_Z[\ell(x, Z)] + \psi(B^\top x)$
$\Rightarrow$ 拡張ラグランジアン: $\mathbb{E}_Z[\ell(x, Z)] + \psi(y) + \lambda^\top(y - B^\top x) + \frac{\rho}{2}\|y - B^\top x\|^2$.

通常の SGD: $x_{t+1} = \arg\min_x \left\{ \langle g_t, x \rangle + \tilde{\psi}(x) + \frac{1}{2\eta_t}\|x - x_t\|^2 \right\} \quad (g_t \in \partial_x \ell(x_t, z_t))$.

SGD-ADMM

$$
\begin{aligned}
x_{t+1} =& \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ g_t^\top x - \lambda_t^\top (B^\top x - y_t) \right. \\
& \left. + \frac{\rho}{2}\|B^\top x - y_t\|^2 + \frac{1}{2\eta_t}\|x - x_t\|_{G_t}^2 \right\}, \\
y_{t+1} =& \operatorname*{argmin}_{y \in \mathcal{Y}} \left\{ \psi(y) - \lambda_t^\top (B^\top x_{t+1} - y) + \frac{\rho}{2}\|B^\top x_{t+1} - y\|^2 \right\} \\
\lambda_{t+1} =& \lambda_t - \rho(B^\top x_{t+1} - y_{t+1}).
\end{aligned}
$$

- y_{t+1} と λ_{t+1} の更新は通常の ADMM と同じ.
- G_t は任意の正定値対称行列.

SGD-ADMM

$\min_x \mathbb{E}_Z[\ell(x, Z)] + \psi(B^\top x)$
$\Rightarrow$ 拡張ラグランジアン: $\mathbb{E}_Z[\ell(x, Z)] + \psi(y) + \lambda^\top(y - B^\top x) + \frac{\rho}{2}\|y - B^\top x\|^2$.

通常の SGD: $x_{t+1} = \arg\min_x \left\{ \langle g_t, x \rangle + \tilde{\psi}(x) + \frac{1}{2\eta_t}\|x - x_t\|^2 \right\} \quad (g_t \in \partial_x \ell(x_t, z_t))$.

SGD-ADMM

$$
\begin{aligned}
x_{t+1} =& \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ g_t^\top x - \lambda_t^\top (B^\top x - y_t) \right. \\
& \qquad \left. + \frac{\rho}{2}\|B^\top x - y_t\|^2 + \frac{1}{2\eta_t}\|x - x_t\|_{G_t}^2 \right\}, \\
y_{t+1} =& \operatorname*{argmin}_{y \in \mathcal{Y}} \left\{ \psi(y) - \lambda_t^\top (B^\top x_{t+1} - y) + \frac{\rho}{2}\|B^\top x_{t+1} - y\|^2 \right\} \\
=& \operatorname{prox}(B^\top x_{t+1} - \lambda_t/\rho | \psi), \\
\lambda_{t+1} =& \lambda_t - \rho(B^\top x_{t+1} - y_{t+1}).
\end{aligned}
$$

- y_{t+1} と λ_{t+1} の更新は通常の ADMM と同じ.
- G_t は任意の正定値対称行列.

RDA-ADMM

通常の RDA: $w_{t+1} = \arg\min_w \left\{ \langle \bar{g}_t, w \rangle + \tilde{\psi}(w) + \frac{1}{2\eta_t} \|w\|^2 \right\} \quad (\bar{g}_t = \frac{1}{t}(g_1 + \cdots + g_t))$

RDA-ADMM

Let $\bar{x}_t = \frac{1}{t}\sum_{\tau=1}^t x_\tau$, $\bar{\lambda}_t = \frac{1}{t}\sum_{\tau=1}^t \lambda_\tau$, $\bar{y}_t = \frac{1}{t}\sum_{\tau=1}^t y_\tau$, $\bar{g}_t = \frac{1}{t}\sum_{\tau=1}^t g_\tau$.

$$
\begin{aligned}
x_{t+1} =& \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \bar{g}_t^\top x - (B\bar{\lambda}_t)^\top x + \frac{\rho}{2t}\|B^\top x\|^2 \right. \\
& \left. + \rho(B^\top \bar{x}_t - \bar{y}_t)^\top B^\top x + \frac{1}{2\eta_t}\|x\|_{G_t}^2 \right\}, \\
y_{t+1} =& \operatorname{prox}(B^\top x_{t+1} - \lambda_t/\rho | \psi), \\
\lambda_{t+1} =& \lambda_t - \rho(B^\top x_{t+1} - y_{t+1}).
\end{aligned}
$$

y_{t+1} と λ_{t+1} の更新は通常の ADMM と同じ.

Convergence analysis

We bound the expected risk:

- Expected risk

$$P(x) = \mathbb{E}_Z[\ell(Z, x)] + \tilde{\psi}(x).$$

Assumptions:

(A1) $\exists G$ s.t. $\forall g \in \partial_x \ell(z, x)$ satisfies $\|g\| \leq G$ for all z, x.

(A2) $\exists L$ s.t. $\forall g \in \partial \psi(y)$ satisfies $\|g\| \leq L$ for all y.

(A3) $\exists R$ s.t. $\forall x \in \mathcal{X}$ satisfies $\|x\| \leq R$.

Convergence rate: bounded gradient

(A1) $\exists G$ s.t. $\forall g \in \partial_x \ell(z, x)$ satisfies $\|g\| \leq G$ for all z, x.

(A2) $\exists L$ s.t. $\forall g \in \partial \psi(y)$ satisfies $\|g\| \leq L$ for all y.

(A3) $\exists R$ s.t. $\forall x \in \mathcal{X}$ satisfies $\|x\| \leq R$.

Theorem (Convergence rate of RDA-ADMM)

Under (A1), (A2), ~~(A3)~~, we have

$$\mathbb{E}_{z_{1:T-1}}[P(\bar{x}_T) - P(x^*)] \leq \frac{1}{T} \sum_{t=2}^{T} \frac{\eta_{t-1}}{2(t-1)} G^2 + \frac{\gamma}{\eta_T} \|x^*\|^2 + \frac{K}{T}.$$

Theorem (Convergence rate of SGD-ADMM)

Under (A1), (A2), (A3), we have

$$\mathbb{E}_{z_{1:T-1}}[P(\bar{x}_T) - P(x^*)] \leq \frac{1}{2T} \sum_{t=2}^{T} \max\left\{ \frac{\gamma}{\eta_t} - \frac{\gamma}{\eta_{t-1}}, 0 \right\} R^2 + \frac{1}{T} \sum_{t=1}^{T} \frac{\eta_t}{2} G^2 + \frac{K}{T}.$$

Both methods have convergence rate $O\left(\frac{1}{\sqrt{T}}\right)$ by letting $\eta_t = \eta_0 \sqrt{t}$ for RDA-ADMM and $\eta_t = \eta_0 / \sqrt{t}$ for SGD-ADMM.

有限和の問題

正則化あり訓練誤差の双対問題

$A = [a_1, a_2, \ldots, a_n] \in \mathbb{R}^{p \times n}$.

$$\min_w \left\{ \frac{1}{n} \sum_{i=1}^{n} f_i(a_i^\top w) + \psi(B^\top w) \right\} \qquad \text{(P: 主)}$$

$$= - \min_{x \in \mathbb{R}^n, y \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^{n} f_i^*(x_i) + \psi^* \left(\frac{y}{n} \right) \mid Ax + By = 0 \right\} \qquad \text{(D: 双対)}$$

最適性条件:

$$a_i^\top w^* \in \nabla f_i^*(x_i^*), \ \frac{1}{n} y^* \in \nabla \psi(u)|_{u = B^\top w^*}, \ Ax^* + By^* = 0.$$

★ 各座標 x_i は各観測値 a_i に対応.

SDCA-ADMM

拡張ラグランジアン:
$\mathcal{L}(x, y, w) := \sum_{i=1}^{n} f_i^*(x_i) + n\psi^*(y/n) - \langle w, Ax + By\rangle + \frac{\rho}{2}\|Ax + By\|^2$.

SDCA-ADMM

For each $t = 1, 2, \ldots$

Choose $i \in \{1, \ldots, n\}$ uniformly at random, and update

$$y^{(t)} \leftarrow \arg\min_{y} \left\{ \mathcal{L}(x^{(t-1)}, y, w^{(t-1)}) + \frac{1}{2}\|y - y^{(t-1)}\|_Q^2 \right\}$$

$$x_i^{(t)} \leftarrow \arg\min_{x_i \in \mathbb{R}} \left\{ \mathcal{L}([x_i; x_{\setminus i}^{(t-1)}], y^{(t)}, w^{(t-1)}) + \frac{1}{2}\|x_i - x_i^{(t-1)}\|_{G_{i,i}}^2 \right\}$$

$$w^{(t)} \leftarrow w^{(t-1)} - \xi\rho\{n(Ax^{(t)} + By^{(t)}) - (n-1)(Ax^{(t-1)} + By^{(t-1)})\}.$$

Q, $G_{i,i}$ はある条件を満たす正定値対称行列.

- 各更新で i-番目の座標 x_i のみ更新.
- w の更新は気を付ける必要がある.

Outline

1. 確率的最適化のより高度な話題
   - 非凸関数の確率的最適化
   - 構造的正則化の最適化

2. 無限次元の確率的最適化：カーネル法
   - 再生核ヒルベルト空間の定義
   - 再生核ヒルベルト空間における最適化

再生核ヒルベルト空間上での最適化
(後の Neural Tangent Kernel ともつながるので紹介)

線形回帰

デザイン行列 $X = (X_{ij}) \in \mathbb{R}^{n \times p}$. $\quad Y = [y_1, \ldots, y_n]^\top \in \mathbb{R}^n$.
真のベクトル $\beta^* \in \mathbb{R}^p$:

$$\text{モデル}: \quad Y = X\beta^* + \boldsymbol{\xi}.$$

リッジ回帰（**Tsykonov** 正則化）

$$\hat{\beta} \leftarrow \underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{1}{n} \|X\beta - Y\|_2^2 + \lambda_n \|\beta\|_2^2.$$

線形回帰

デザイン行列 $X = (X_{ij}) \in \mathbb{R}^{n \times p}$. $\quad Y = [y_1, \ldots, y_n]^\top \in \mathbb{R}^n$.
真のベクトル $\beta^* \in \mathbb{R}^p$:

$$\text{モデル}: \quad Y = X\beta^* + \boldsymbol{\xi}.$$

リッジ回帰（**Tsykonov** 正則化）

$$\hat{\beta} \leftarrow \underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{1}{n} \|X\beta - Y\|_2^2 + \lambda_n \|\beta\|_2^2.$$

変数変換:

- 正則化項のため，$\hat{\beta} \in \mathrm{Ker}(X)^\perp$．つまり，$\hat{\beta} \in \mathrm{Im}(X^\top)$.
- ある $\hat{\alpha} \in \mathbb{R}^n$ が存在して，$\hat{\beta} = X^\top \hat{\alpha}$ と書ける.

$$\text{(等価な問題)} \qquad \hat{\alpha} \leftarrow \underset{\alpha \in \mathbb{R}^n}{\arg\min} \frac{1}{n} \|XX^\top \alpha - Y\|_2^2 + \lambda_n \alpha^\top (XX^\top) \alpha.$$

※ $(XX^\top)_{ij} = x_i^\top x_j$ より，観測値 x_i と x_j の内積さえ計算できればよい.

リッジ回帰のカーネル化

リッジ回帰（変数変換版）

$$\hat{\alpha} \leftarrow \underset{\alpha \in \mathbb{R}^n}{\arg\min} \frac{1}{n} \|(XX^\top)\alpha - Y\|_2^2 + \lambda_n \alpha^\top (XX^\top)\alpha.$$

※ $(XX^\top)_{ij} = x_i^\top x_j$ はサンプル x_i と x_j の内積.

リッジ回帰のカーネル化

リッジ回帰（変数変換版）

$$\hat{\alpha} \leftarrow \underset{\alpha \in \mathbb{R}^n}{\arg\min} \frac{1}{n} \|(XX^\top)\alpha - Y\|_2^2 + \lambda_n \alpha^\top (XX^\top)\alpha.$$

※ $(XX^\top)_{ij} = x_i^\top x_j$ はサンプル x_i と x_j の内積.

- カーネル法のアイディア
  x の間の内積を他の非線形な関数で置き換える:

$$x_i^\top x_j \ \rightarrow \ k(x_i, x_j).$$

この $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ を**カーネル関数**と呼ぶ.

カーネル関数の満たすべき条件

- 対称性: $k(x, x') = k(x', x)$.
- 正値性: $\sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j k(x_i, x_j) \geq 0, \quad (\forall \{x_i\}_{i=1}^m, \ \{\alpha_i\}_{i=1}^m, \ m)$.

逆にこの性質を満たす関数なら何でもカーネル法で用いて良い.

カーネルリッジ回帰

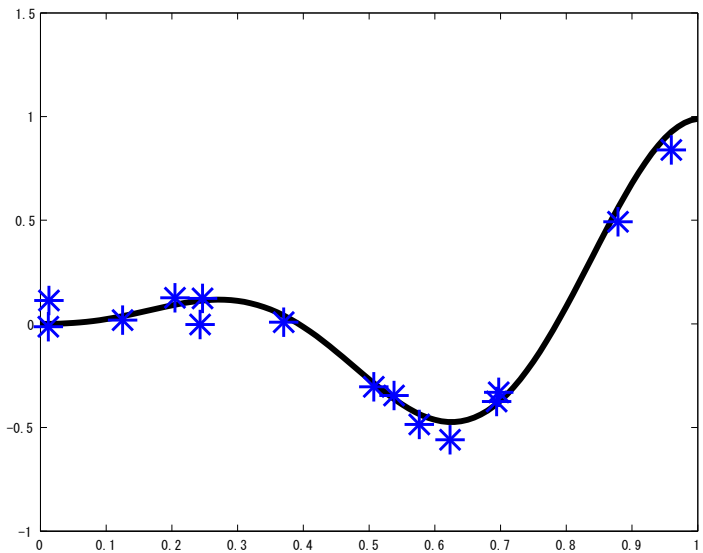
カーネルリッジ回帰: $K = (k(x_i, x_j))_{i,j=1}^n$ として,

$$\hat{\alpha} \leftarrow \underset{\beta \in \mathbb{R}^n}{\arg\min} \frac{1}{n} \|K\alpha - Y\|_2^2 + \lambda_n \alpha^\top K \alpha.$$

新しい入力 x に対しては,

$$y = \sum_{i=1}^n k(x, x_i)\hat{\alpha}_i$$

で予測.



カーネルリッジ回帰

カーネルリッジ回帰: $K = (k(x_i, x_j))_{i,j=1}^n$ として,

$$\hat{\alpha} \leftarrow \underset{\beta \in \mathbb{R}^n}{\arg\min} \frac{1}{n} \|K\alpha - Y\|_2^2 + \lambda_n \alpha^\top K \alpha.$$

新しい入力 x に対しては,

$$y = \sum_{i=1}^n k(x, x_i)\hat{\alpha}_i$$

で予測.

カーネル関数 $\Leftrightarrow$ 再生核ヒルベルト空間 **(RKHS)**

$k(x, x')$ $\qquad$ $\mathcal{H}_k$

ある $\phi(x) : \mathbb{R}^p \to \mathcal{H}_k$ が存在して,

- $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}_k}$.
- カーネルトリック: $\langle \sum_{i=1}^n \alpha_i \phi(x_i), \phi(x) \rangle_{\mathcal{H}_k} = \sum_{i=1}^n \alpha_i k(x_i, x)$.
  →カーネル関数の値さえ計算できれば良い.

再生核ヒルベルト空間 (Reproducing Kernel Hilbert Space, RKHS)

入力データの分布：P_X，対応する L_2 空間：$L_2(P_X) = \{f \mid \mathbb{E}_{X\sim P_X}[f(X)^2] < \infty\}$.
カーネル関数は以下のように分解できる (Steinwart and Scovel, 2012):

$$k(x, x') = \sum_{j=1}^{\infty} \mu_j e_j(x) e_j(x').$$

- $(e_j)_{j=1}^{\infty}$ は $L_2(P_X)$ 内の正規直交基底: $\|e_j\|_{L_2(P_X)} = 1$, $\langle e_j, e_{j'}\rangle_{L_2(P_X)} = 0$ $(j \neq j')$.
- $\mu_j \geq 0$.

Definition (再生核ヒルベルト空間 ($\mathcal{H}_k$))

- $\langle f, g\rangle_{\mathcal{H}_k} := \sum_{j=1}^{\infty} \frac{1}{\mu_j}\alpha_j\beta_j$ for $f = \sum_{j=1}^{\infty}\alpha_j e_j$, $g = \sum_{j=1}^{\infty}\beta_j e_j \in L_2(P_X)$.
- $\|f\|_{\mathcal{H}_k} := \sqrt{\langle f, f\rangle_{\mathcal{H}_k}}$.
- $\mathcal{H}_k := \{f \in L_2(P_X) \mid \|f\|_{\mathcal{H}_k} < \infty\}$ equipped with $\langle\cdot,\cdot\rangle_{\mathcal{H}_k}$.

再生性: $f \in \mathcal{H}_k$ に対して $f(x)$ は内積の形で「再生」される:

$$f(x) = \langle f, k(x, \cdot)\rangle_{\mathcal{H}_k}.$$

再生核ヒルベルト空間の性質

$$\phi_k(x) = k(x, \cdot) \in \mathcal{H}_k$$

と書けば，$k(x, x') = \langle \phi_k(x), \phi_k(x') \rangle_{\mathcal{H}_k}$ と書ける．この ϕ_k を特徴写像とも言う．

カーネル関数に対応する積分作用素 $T_k : L_2(P_X) \to L_2(P_X)$:

$$T_k f := \int f(x) k(x, \cdot) \mathrm{d}P_X(x).$$

- 先のカーネル関数の分解は T_k のスペクトル分解に対応．
- 再生核ヒルベルト空間 $\mathcal{H}_k$ は以下のようにも書ける:

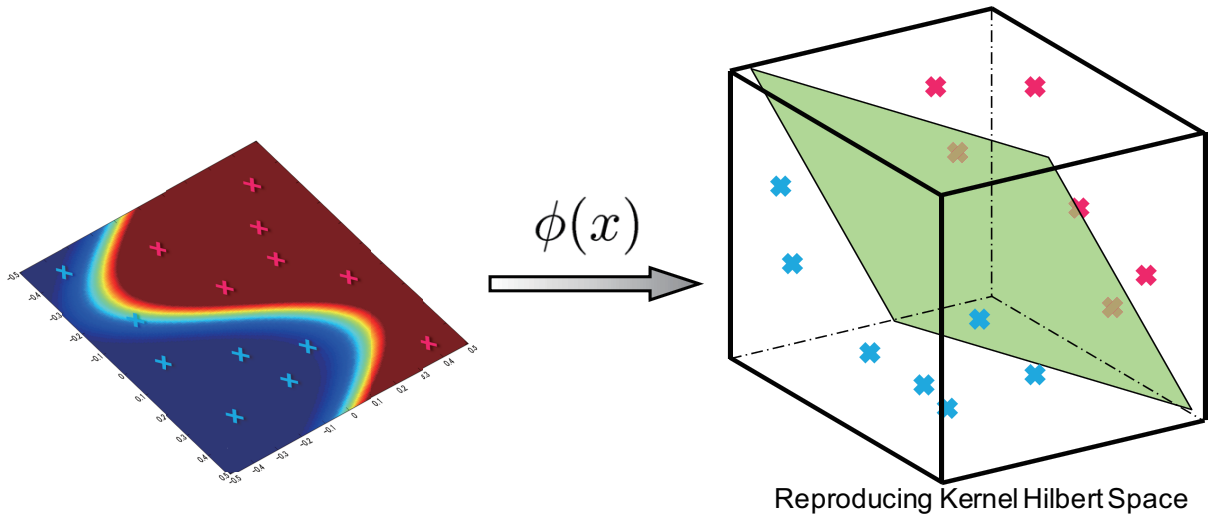
$$\mathcal{H}_k = T_k^{1/2} L_2(P_X).$$

- $\|f\|_{\mathcal{H}_k} = \inf\{\|h\|_{L_2(P_X)} \mid f = T_k^{1/2} h,\ h \in L_2(P_X)\}$.
  - $f \in \mathcal{H}_k$ は $f(x) = \sum_{j=1}^{\infty} a_j \sqrt{\mu_j} e_j(x)$ と書けて，$\|f\|_{\mathcal{H}_k} = \sqrt{\sum_{j=1}^{\infty} a_j^2}$.
  - $(e_j)_j$ は L_2 内の正規直交基底，$(\sqrt{\mu_j} e_j)_j$ は RKHS 内の完全正規直交基底．
  - 特徴写像 $\phi_k(x) = k(x, \cdot) \in \mathcal{H}_k$ を完全正規直交基底に関する係数で表現すると

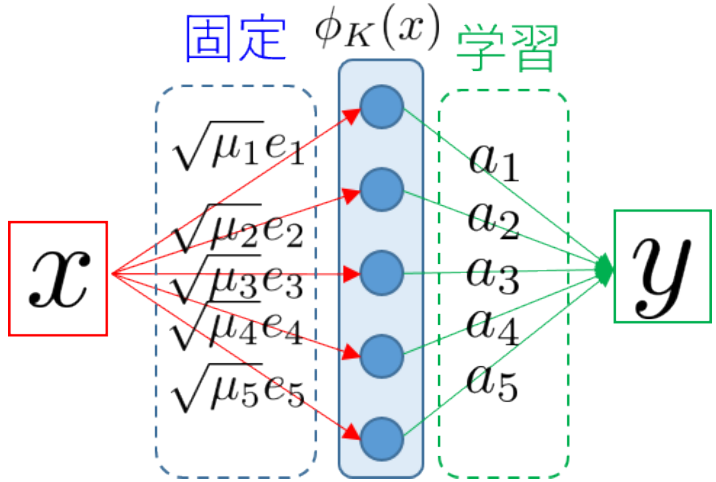
$$\phi_k(x) = (\sqrt{\mu_1} e_1(x), \sqrt{\mu_2} e_2(x), \dots)^\top$$

再生核ヒルベルト空間のイメージ

- 非線形な推論を再生核ヒルベルト空間への非線形写像 ϕ を用いて行う.
- 再生核ヒルベルト空間では線形な処理をする.



- カーネル法は第一層を固定し第二層目のパラメータを学習する横幅無限大の 2 層ニューラルネットワークともみなせる.
(“浅い”学習手法の代表例)



カーネルリッジ回帰の再定式化

- 再生性: $f \in \mathcal{H}_k$ に対し
$$f(x) = \langle f, \phi(x) \rangle_{\mathcal{H}_k}.$$
- カーネルリッジ回帰の再定式化

$$\hat{f} \leftarrow \min_{f \in \mathcal{H}_k} \quad \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + C \|f\|_{\mathcal{H}_k}^2$$

- 表現定理

$$\exists \alpha_i \in \mathbb{R} \quad \text{s.t.} \quad \hat{f}(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x),$$

$$\Rightarrow \quad \|\hat{f}\|_{\mathcal{H}_k} = \sqrt{\textstyle\sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j)} = \sqrt{\boldsymbol{\alpha}^\top K \boldsymbol{\alpha}}.$$

さきほどのカーネルリッジ回帰の定式化と一致.

カーネルの例

- ガウシアンカーネル

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$$

- 多項式カーネル

$$k(x, x') = \left(1 + x^\top x'\right)^p$$

- χ^2-カーネル

$$k(x, x') = \exp\left(-\gamma^2 \sum_{j=1}^{d} \frac{(x_j - x'_j)^2}{(x_j + x'_j)}\right)$$

- Matérn-kernel

$$k(x, x') = \int_{\mathbb{R}^d} e^{\mathrm{i}\lambda^\top (x - x')} \frac{1}{(1 + \|\lambda\|^2)^{\alpha + d/2}} \mathrm{d}\lambda$$

- グラフカーネル，時系列カーネル，...

再生核ヒルベルト空間内の確率的最適化

問題設定:

$$y_i = f^{\mathrm{o}}(x_i) + \xi_i.$$

$(x_i, y_i)_{i=1}^n$ から f^{o} を推定したい．(f^{o} は $\mathcal{H}_k$ にほぼ入っている)

期待損失の変形:

$$\mathbb{E}[(f(X) - Y)^2] = \mathbb{E}[(f(X) - f^{\mathrm{o}}(X) - \xi)^2] = \mathbb{E}[(f(X) - f^{\mathrm{o}}(X))^2] + \sigma^2$$

$\to \min_{f \in \mathcal{H}_k} \mathbb{E}[(f(X) - Y)^2]$ を解けば f^{o} が求まる．

$K_x = k(x, \cdot) \in \mathcal{H}_k$ とすると，$f(x) = \langle f, K_x \rangle_{\mathcal{H}_k}$ より $L(f) = \mathbb{E}[(f(X) - Y)^2]$ の RKHS 内での Frechet 微分は以下の通り:

$$\nabla L(f) = 2\mathbb{E}[K_X(\langle K_X, f \rangle_{\mathcal{H}_k} - Y)] = 2(\underbrace{\mathbb{E}[K_X K_X^*]}_{=:\Sigma} f - \mathbb{E}[K_X Y]) = 2(\Sigma f - \mathbb{E}[K_X Y]).$$

- 期待損失の勾配法:
$$f_t^* = f_{t-1}^* - \eta 2(\Sigma f_{t-1}^* - \mathbb{E}[K_X Y]).$$
- 経験損失の勾配法 ($\widehat{\mathbb{E}}[\cdot]$ は標本平均):
$$\hat{f}_t = \hat{f}_{t-1} - \eta 2(\widehat{\Sigma} \hat{f}_{t-1} - \widehat{\mathbb{E}}[K_X Y]).$$
- 確率的勾配による更新:
$$g_t = g_{t-1} - \eta 2(K_{x_{i_t}} K_{x_{i_t}}^* g_{t-1} - K_{x_{i_t}} y_{i_t}).$$

※ $(x_{i_t}, y_{i_t})_{t=1}^{\infty}$ は $(x_i, y_i)_{i=1}^n$ から i.i.d. 一様に取得．

勾配のスムージングとしての見方

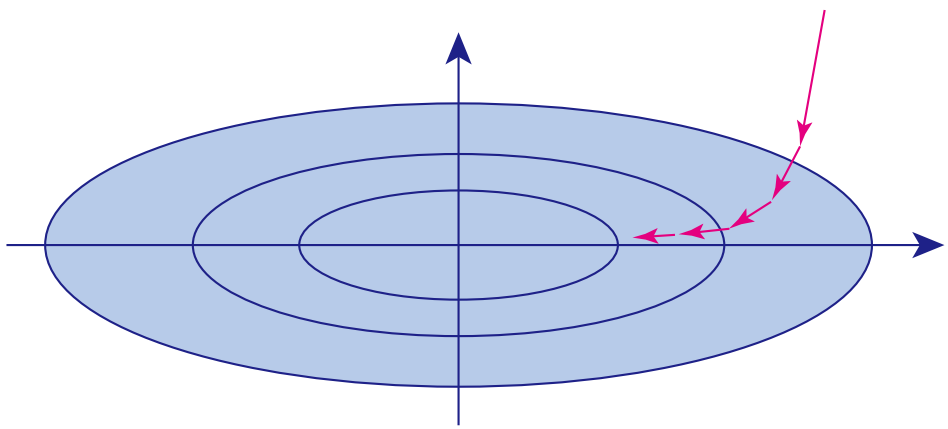
関数値の更新式:

$$f_t^*(x) = f_{t-1}^*(x) - 2\eta \int k(x, X) \underbrace{(f_{t-1}^*(X) - Y)}_{\to f_{t-1}^*(X) - f^{\circ}(X)} \mathrm{d}P(X, Y)$$

$$= f_{t-1}^*(x) - 2\eta T_k(f_{t-1}^* - f^{\circ})(x).$$

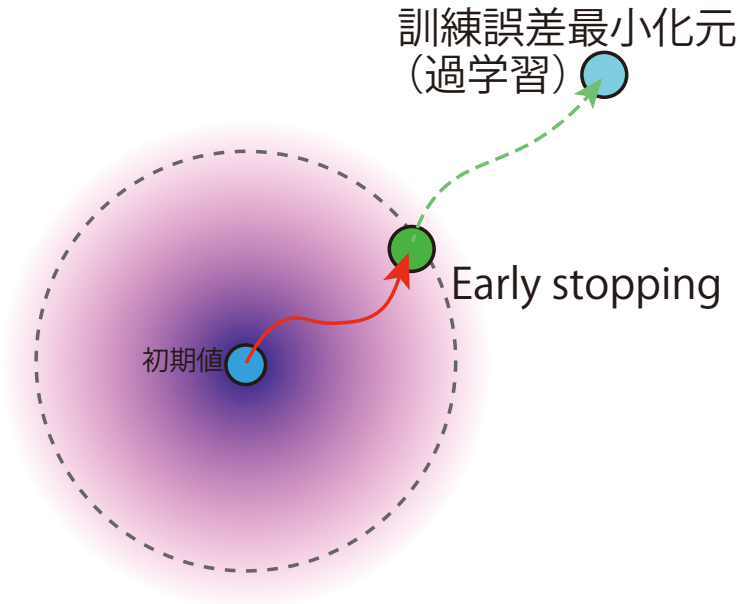
積分作用素 T_k は高周波成分を抑制する作用がある.

- RKHS 内の勾配は L_2 内の関数勾配を T_k によって平滑化したものになっている. (実際は T_k のサンプルからの推定値を使う)
- 高周波成分が出てくる前に止めれば過学習を防げる.
  → **Early stopping**
- 迂闊に Newton 法などを使うと危険.



Early stopping による正則化

Early stopping による正則化



バイアス-バリアンス分解

$$\underbrace{\|f^{\mathrm{o}} - \hat{f}\|_{L_2(P_X)}}_{\text{Estimation error}} \leq \underbrace{\|f^{\mathrm{o}} - \check{f}\|_{L_2(P_X)}}_{\substack{\text{Approximation error} \\ \text{(bias)}}} + \underbrace{\|\check{f} - \hat{f}\|_{L_2(P_X)}}_{\substack{\text{Sample deviation} \\ \text{(variance)}}}$$

訓練誤差最小化元に達する前に止める (early stopping) ことで正則化が働く.
無限次元モデル (RKHS) は過学習しやすいので気を付ける必要がある.

解析に用いる条件

通常，以下の条件を考える．（統計理論でも同様の仮定を課す定番の仮定）
(Caponnetto and de Vito, 2007, Dieuleveut et al., 2016, Pillaud-Vivien et al., 2018)

- $\mu_i = O(i^{-\alpha})$ for $\alpha > 1$.
  α は RKHS$\mathcal{H}_k$ の複雑さを特徴づける．(小さい α: 複雑，大きい α: 単純)
- $f^{\mathrm{o}} \in T^r(L_2(P_X))$ for $r > 0$.
  f^{o} が RKHS からどれだけ “はみ出ているか” を特徴づけ．
  $r = 1/2$ は $f^{\mathrm{o}} \in \mathcal{H}_k$ に対応．($r < 1/2$: はみ出てる, $r \geq 1/2$: 含まれる)
- $\|f\|_{L_\infty(P_X)} \lesssim \|f\|_{L_2(P_X)}^{1-\mu} \|f\|_{\mathcal{H}_k}^{\mu}$ $(\forall f \in \mathcal{H}_k)$ for $\mu \in (0, 1]$.
  $\mathcal{H}_k$ に含まれている関数の滑らかさを特徴づけ．（小さい μ: 滑らか）

※ 最後の条件について: $f \in W^m([0,1]^d)$ (Sobolev 空間) かつ P_X の台が $[0,1]^d$ で密度関数を持ち，その密度が下からある定数 $c > 0$ で抑えられていれば，$\mu = d/(2m)$ でなりたつ．

収束レート

バイアス-バリアンスの分解:

$$\|f^{\mathrm{o}} - g_t\|_{L_2(P_X)}^2 \lesssim \underbrace{\|f^{\mathrm{o}} - f_t^*\|_{L_2(P_X)}^2}_{(a)\text{: Bias}} + \underbrace{\|f_t^* - \hat{f}_t\|_{L_2(P_X)}^2}_{(b)\text{: Variance}} + \underbrace{\|\hat{f}_t - g_t\|_{L_2(P_X)}^2}_{(c)\text{: SGD deviation}}$$

$$\text{(a) } (\eta t)^{-2r}, \quad \text{(b) } \frac{(\eta t)^{1/\alpha} + (\eta t)^{\mu - 2r}}{n}, \quad \text{(c) } \eta (\eta t)^{1/\alpha - 1}$$

(a) 勾配法の解のデータに関する期待値と真の関数とのズレ (Bias).
(b) 勾配法の解の分散 (Variance).
(c) 確率的勾配を用いることによる変動.

更新数 t を大きくすると Bias は減るが Variance が増える．これらをバランスする必要がある (Early stopping).

Theorem (Multi-pass SGD の収束レート (Pillaud-Vivien et al., 2018))

$\eta = 1/(4 \sup_x k(x,x)^2)$ とする.

- $\mu\alpha < 2r\alpha + 1 < \alpha$ の時，$t = \Theta(n^{\alpha/(2r\alpha+1)})$ とすれば,

$$\mathbb{E}[L(g_t)] - L(f^{\mathrm{o}}) = O(n^{-2r\alpha/(2r\alpha+1)}).$$

- $\mu\alpha \geq 2r\alpha + 1$ の時，$t = \Theta(n^{\frac{1}{\mu}} (\log n)^{\frac{1}{\mu}})$ とすれば，$\mathbb{E}[L(g_t)] - L(f^{\mathrm{o}}) = O(n^{-2r/\mu})$.

Natural gradient の収束

Natural gradient (自然勾配法):

$$\hat{f}_t = \hat{f}_{t-1} - \eta(\Sigma + \lambda \mathrm{I})^{-1}(\widehat{\Sigma}\hat{f}_{t-1} - \widehat{\mathbb{E}}[K_X Y]).$$

(unlabeled data が沢山あり Σ は良く推定できる設定; GD の解析 (Murata and Suzuki, 2020))

Theorem (Natural gradient の収束 (Amari et al., 2020))

$$\mathbb{E}[\|\hat{f}_t - f^{\mathrm{o}}\|^2_{L_2(P_X)}] \lesssim B(t) + V(t),$$

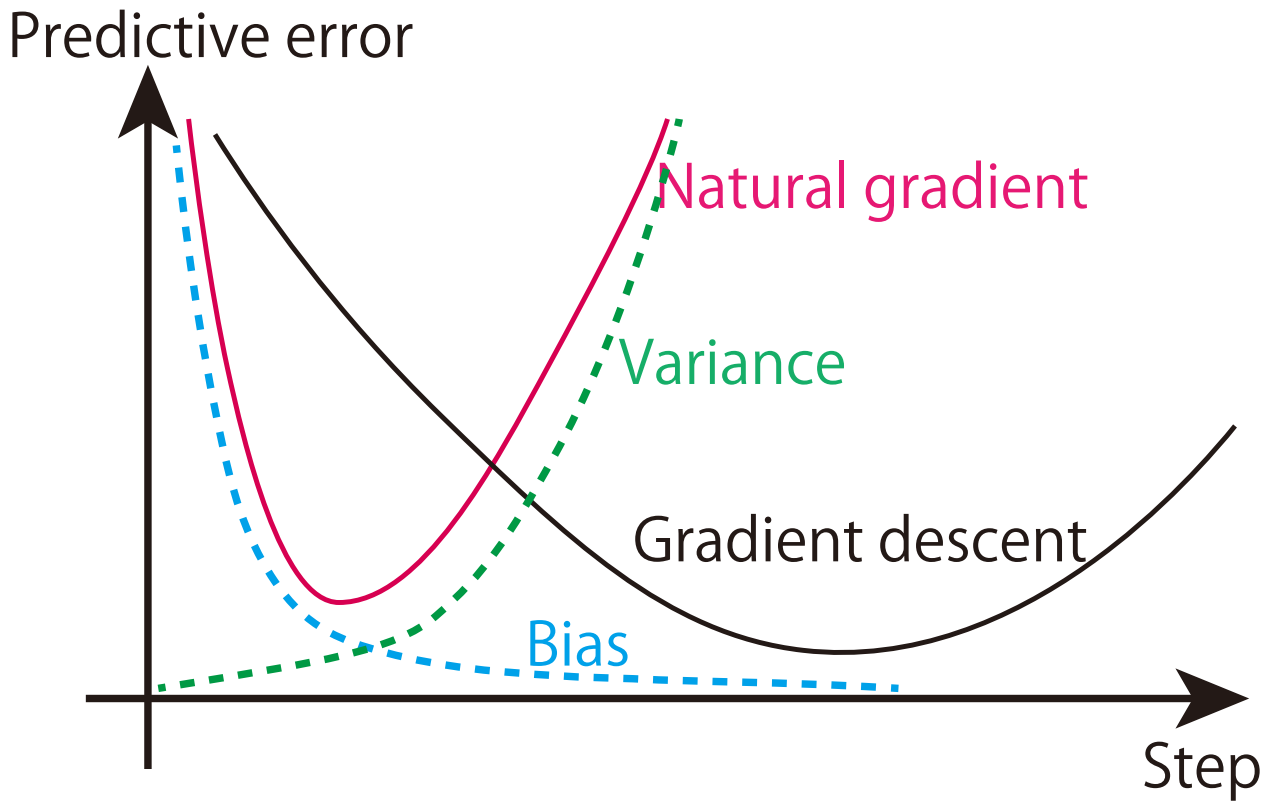
ただし, $B(t) = \exp(-\eta t) \vee (\lambda/(\eta t))^{2r}$,

$$V(t) = (1+\eta t)\frac{\lambda^{-1}B(t) + \lambda^{-\frac{1}{\alpha}}}{n} + (1+t\eta)^4 \frac{(1 \vee \lambda^{2r-\mu})\lambda^{-\frac{1}{\alpha}}}{n}.$$

特に, $\lambda = n^{-\frac{\alpha}{2r\alpha+1}}$, $t = \Theta(\log(n))$ で $\mathbb{E}[\|\hat{f}_t - f^{\mathrm{o}}\|^2_{L_2(P_X)}] = O(n^{-\frac{2r\alpha}{2r\alpha+1}}\log(n)^4)$.

※ バイアスは急速に収束するが, バリアンスも速く増大する.
→ Preconditioning のため高周波成分が早めに出現する. より早めに止めないと過学習する.

収束の様子


Predictive error
Natural gradient
Variance
Gradient descent
Bias
Step

作用素 Bernstein の不等式

- $\Sigma = \mathbb{E}_x[K_x K_x^*]$: $\Sigma f = \int k(\cdot, x) f(x) \mathrm{d}P_x(x)$
- $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n K_{x_i} K_{x_i}^*$: $\widehat{\Sigma} f = \frac{1}{n}\sum_{i=1}^n k(\cdot, x_i) f(x_i)$

$\Sigma_\lambda := \Sigma + \lambda \mathrm{I}$, $\mathcal{F}_\infty(\lambda) := \sup_x K_x^* \Sigma_\lambda^{-1} K_x$ とする．以下のような評価が必要:

$$\|\Sigma_\lambda^{-1}(\Sigma - \widehat{\Sigma})\Sigma_\lambda^{-1}\| \lesssim \sqrt{\frac{\mathcal{F}_\infty(\lambda)\beta}{n}} + \frac{(1 + \mathcal{F}_\infty(\lambda))\beta}{n}$$

with prob. $1 - \delta$．ただし，$\beta = \log(\frac{4\mathrm{Tr}[\Sigma\Sigma_\lambda^{-1}]}{\delta})$.
→ 経験分布と真の分布のずれをバウンド.

Theorem (自己共役作用素の Bernstein の不等式 (Minsker, 2017))

$(X_i)_{i=1}^n$ は独立な自己共役作用素の確率変数で $\mathbb{E}[X_i] = 0$ かつ，
$\sigma^2 \geq \|\sum_{i=1}^n \mathbb{E}[X_i^2]\|$, $U \geq \|X_i\|$ とする．$r(A) = \mathrm{Tr}[A]/\|A\|$ として，

$$P\left(\left\|\sum_{i=1}^n X_i\right\| \geq t\right) \leq 14 r(\textstyle\sum_{i=1}^n \mathbb{E}[X_i^2]) \exp\left(-\frac{t^2}{2(\sigma^2 + tU/3)}\right).$$

$X_i = \Sigma_\lambda^{-1} K_{x_i} K_{x_i}^* \Sigma_\lambda^{-1}$ とする． (Tropp (2012) も参照)

正則化ありの確率的最適化

二乗損失を拡張して，一般の滑らかな凸損失関数 ℓ を考える．(判別問題など)
正則化ありの期待損失最小化:

$$\min_{f \in \mathcal{H}_k} \quad \mathbb{E}[\ell(Y, f(X))] + \lambda \|f\|_{\mathcal{H}_k}^2 =: L_\lambda(f).$$

これを SGD で解く．目的関数が λ-強凸であることを利用．

$$g_{t+1} = g_t - \eta_t(\ell'(y_t, g_t(x_t)) + \lambda g_t).$$

$$\bar{g}_{T+1} = \sum_{t=1}^{T+1} \frac{2(c_0+t-1)}{(2c_0+T)(T+1)} g_t \quad \text{(多項式平均)}.$$

仮定：(i) ℓ は γ-平滑，$\|\ell'\|_\infty \le M$, (ii) $k(x,x) \le 1$.　　$g_\lambda = \mathrm{argmin}_{g \in \mathcal{H}_k} L_\lambda(g)$.

Theorem (Nitanda and Suzuki (2019))

適切な $c_0 > 0$ に対して $\eta_t = 2/(\lambda(c_0 + t))$ とすれば，

$$\mathbb{E}[L_\lambda(\bar{g}_{T+1}) - L_\lambda(g_\lambda)] \lesssim \frac{M^2}{\lambda(c_0+T)} + \frac{\gamma+\lambda}{T+1}\|g_1 - g_\lambda\|_{\mathcal{H}_k}^2.$$

さらにマルチンゲール確率集中不等式より High probability bound も得られる．
判別問題なら strong low noise condition のもと判別誤差の**指数収束**も示せる．

マルチンゲール Hoeffding の不等式

Theorem (マルチンゲール Hoeffding 型集中不等式 (Pinelis, 1994))

確率変数列: $D_1, \ldots, D_T \in \mathcal{H}_k$. $\mathbb{E}[D_t] = 0$, $\|D_t\|_{\mathcal{H}_k} \leq R_t$ *(a.s.)* とする.
$\forall \epsilon > 0$ に対し

$$P\left[\max_{1 \leq t \leq T} \|\sum_{s=1}^{t} D_s\|_{\mathcal{H}_k} \geq \epsilon\right] \leq 2\exp\left(-\frac{\epsilon^2}{2\sum_{t=1}^{T} R_t^2}\right).$$

$$D_t = \mathbb{E}[\bar{g}_{T+1}|Z_1, \ldots, Z_t] - \mathbb{E}[\bar{g}_{T+1}|Z_1, \ldots, Z_{t-1}],$$

ただし $Z_t = (x_t, y_t)$ とすれば, $\sum_{t=1}^{T} D_t = \bar{g}_{T+1} - \mathbb{E}[\bar{g}_{T+1}]$ となり, 期待値と実現値のずれを抑えられる.

マルチンゲール Hoeffding の不等式

Theorem (マルチンゲール Hoeffding 型集中不等式 (Pinelis, 1994))

確率変数列: $D_1, \ldots, D_T \in \mathcal{H}_k$. $\mathbb{E}[D_t] = 0$, $\|D_t\|_{\mathcal{H}_k} \leq R_t$ *(a.s.)* とする.
$\forall \epsilon > 0$ に対し

$$P\left[\max_{1 \leq t \leq T} \|\sum_{s=1}^{t} D_s\|_{\mathcal{H}_k} \geq \epsilon\right] \leq 2\exp\left(-\frac{\epsilon^2}{2\sum_{t=1}^{T} R_t^2}\right).$$

$$D_t = \mathbb{E}[\bar{g}_{T+1}|Z_1, \ldots, Z_t] - \mathbb{E}[\bar{g}_{T+1}|Z_1, \ldots, Z_{t-1}],$$

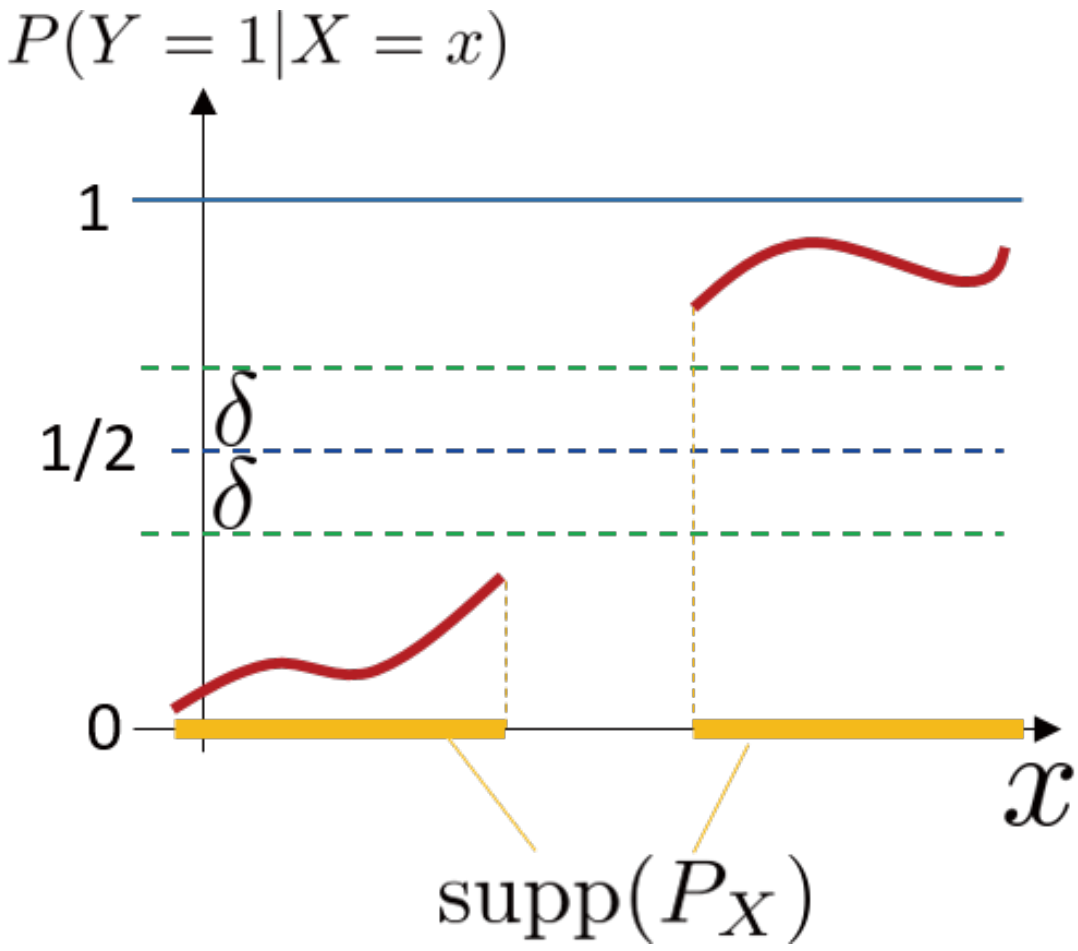
ただし $Z_t = (x_t, y_t)$ とすれば, $\sum_{t=1}^{T} D_t = \bar{g}_{T+1} - \mathbb{E}[\bar{g}_{T+1}]$ となり, 期待値と実現値のずれを抑えられる.

(補足) $\mathcal{L}_\lambda$ は RKHS ノルムに関して λ-強凸であることより,

$$\|\bar{g}_{T+1} - g_\lambda\|_{\mathcal{H}_k} \leq O(\frac{1}{\lambda^2 T})$$

が高い確率で成り立つ. 実は $\|\cdot\|_\infty \leq \|\cdot\|_{\mathcal{H}_k}$ でもあるので,
$|P(Y = 1|X) - P(Y = -1|X)| \geq \delta$ なるマージン条件 (strong low noise condition) のもと, 完全な判別が高い確率でできるようになる.

(参考) Strong low noise condition



Z. Allen-Zhu and E. Hazan. Variance reduction for faster non-convex optimization. arXiv preprint arXiv:1603.05643, 2016.

S. Amari, J. Ba, R. Grosse, X. Li, A. Nitanda, T. Suzuki, D. Wu, and J. Xu. When does preconditioning help or hurt generalization?, 2020.

A. Caponnetto and E. de Vito. Optimal rates for regularized least-squares algorithm. Foundations of Computational Mathematics, 7(3):331–368, 2007.

A. R. Conn, N. I. Gould, and P. L. Toint. Trust region methods, volume 1. Siam, 2000.

W. Deng and W. Yin. On the global and linear convergence of the generalized alternating direction method of multipliers. Technical report, Rice University CAAM TR12-14, 2012.

A. Dieuleveut, F. Bach, et al. Nonparametric stochastic approximation with large step-sizes. The Annals of Statistics, 44(4):1363–1399, 2016.

D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite-element approximations. Computers & Mathematics with Applications, 2:17–40, 1976.

R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In Proceedings of The 28th Conference on Learning Theory, pages 797–842, 2015.

S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. SIAM Journal on Optimization, 23(4):2341–2368, 2013.

B. He and X. Yuan. On the $O(1/n)$ convergence rate of the Douglas-Rachford alternating direction method. SIAM J. Numerical Analisis, 50(2):700–709, 2012.

M. Hestenes. Multiplier and gradient methods. Journal of Optimization Theory & Applications, 4:303–320, 1969.

M. Hong and Z.-Q. Luo. On the linear convergence of the alternating direction method of multipliers. arXiv preprint arXiv:1208.3922, 2012a.

M. Hong and Z.-Q. Luo. On the linear convergence of the alternating direction method of multipliers. Technical report, 2012b. arXiv:1208.3922.

C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan. How to escape saddle points efficiently. In International Conference on Machine Learning, pages 1724–1732, 2017a.

C. Jin, P. Netrapalli, and M. I. Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. arXiv preprint arXiv:1711.10456, 2017b.

Z. Li. Ssrgd: Simple stochastic recursive gradient descent for escaping saddle points. In Advances in Neural Information Processing Systems, pages 1523–1533, 2019.

Z. Li and J. Li. A simple proximal stochastic gradient method for nonsmooth nonconvex optimization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, Advances in Neural Information Processing Systems 31, pages 5564–5574. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/7800-a-simple-proximal-stochastic-gradient-method-for-nonsmooth-n pdf.

Y. Liu, F. Shang, and J. Cheng. Accelerated variance reduced stochastic admm. In Thirty-First AAAI Conference on Artificial Intelligence, pages 2287–2293, 2017.

S. Minsker. On some extensions of Bernstein's inequality for self-adjoint operators. Statistics & Probability Letters, 127:111–119, 2017.

J. F. Mota, J. M. Xavier, P. M. Aguiar, and M. Püschel. A proof of convergence for the alternating direction method of multipliers applied to polyhedral-constrained functions. arXiv preprint arXiv:1112.2295, 2011.

T. Murata and T. Suzuki. Gradient descent in rkhs with importance labeling, 2020.

L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In D. Precup and Y. W. Teh, editors, Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pages 2613–2621, International Convention Centre, Sydney, Australia, 06–11

Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/nguyen17b.html.

A. Nitanda and T. Suzuki. Stochastic gradient descent with exponential convergence rates of expected classification errors. In K. Chaudhuri and M. Sugiyama, editors, Proceedings of Machine Learning Research, volume 89 of Proceedings of Machine Learning Research, pages 1417–1426. PMLR, 16–18 Apr 2019. URL http://proceedings.mlr.press/v89/nitanda19a.html.

H. Ouyang, N. He, L. Q. Tran, and A. Gray. Stochastic alternating direction method of multipliers. In Proceedings of the 30th International Conference on Machine Learning, 2013.

N. H. Pham, L. M. Nguyen, D. T. Phan, and Q. Tran-Dinh. Proxsarah: An efficient algorithmic framework for stochastic composite nonconvex optimization. Journal of Machine Learning Research, 21(110):1–48, 2020.

L. Pillaud-Vivien, A. Rudi, and F. Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. In Advances in Neural Information Processing Systems, pages 8114–8124, 2018.

I. Pinelis. Optimum bounds for the distributions of martingales in banach spaces. The Annals of Probability, pages 1679–1706, 1994.

M. Powell. A method for nonlinear constraints in minimization problems. In R. Fletcher, editor, Optimization, pages 283–298. Academic Press, London, New York, 1969.

S. J. Reddi, A. Hefny, S. Sra, B. Póczós, and A. Smola. Stochastic variance reduction for nonconvex optimization. arXiv preprint arXiv:1603.06160, 2016.

R. T. Rockafellar. Augmented Lagrangians and applications of the proximal point algorithm in convex programming. Mathematics of Operations Research, 1: 97–116, 1976.

I. Steinwart and C. Scovel. Mercer ' s theorem on general domains: on the interaction between measures, kernels, and RKHSs. Constructive Approximation, 35(3):363–417, 2012.

J. Sun, Q. Qu, and J. Wright. When are nonconvex problems not scary? arXiv preprint arXiv:1510.06096, 2015.

T. Suzuki. Dual averaging and proximal gradient descent for online alternating direction multiplier method. In Proceedings of the 30th International Conference on Machine Learning, pages 392–400, 2013.

T. Suzuki. Stochastic dual coordinate ascent with alternating direction method of multipliers. In Proceedings of the 31th International Conference on Machine Learning, pages 736–744, 2014.

J. A. Tropp. User-friendly tools for random matrices: An introduction. Technical report, 2012.

S. Zheng and J. T. Kwok. Fast-and-light stochastic ADMM. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, pages 2407–2413, 2016.