Mirror Descent: from theory to practice

Danya Merkulov

Introduction

Theory vs Practice in 1st order stochastic optimization in NN

Theory

- Optimal 1st order algorithm – mirror descent with rates:
- Non – smooth $\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$
- Smooth $\mathcal{O}\left(\frac{1}{T^2}\right)$

Practice

- Non smooth (even non convex), but usually
- Various variants of SGD are used (Adagrad, Adam, RMSProp, etc.)

Why don't we use an optimal algorithm (MD) for optimization in NN training?

Optimal algorithm?

- Means, that upper bounds for this algorithm meets lower bounds for this class of problems (convex, non-smooth optimization in our case)

Theorem (Nesterov.) Let $\mathcal{B} = \{x \mid \|x - x^0\|_2 \leq D\}$. Assume, $x^* \in \mathcal{B}$. There exists a convex function f in $C_L^0(\mathcal{B})$ (with $L > 0$), such that for $0 \leq k \leq n-1$, the lower-bound

$$f(x^k) - f(x^*) \geq \frac{LD}{2(1+\sqrt{k+1})},$$

holds for **any algorithm** that generates x^k by linearly combining the previous iterates and subgradients.

Projected Subgradient Descent

$$f(\overline{x}) - f^* \leq GR\frac{1}{\sqrt{T}}$$

Mirror Descent

$$f(\overline{x}) - f^* \leq \sqrt{\frac{2MG^2}{T}}$$

(Projected) (Sub)gradient Descent

$$\min_{x \in \mathbb{R}^n} f(x) \qquad x_{k+1} = x_k - \alpha_k g_k$$ (Sub)gradient descent

$$\min_{x \in S} f(x) \qquad x_{k+1} = \Pi_S \left\{ x_k - \alpha_k g_k \right\}$$ Projected subgradient descent

Bounds are usually obtained in a following way:

$$\begin{aligned} \|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\ &= \|x_k - x^*\|^2 + \alpha_k^2 g_k^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \end{aligned}$$

$$2\alpha_k \langle g_k, x_k - x^* \rangle = \|x_k - x^*\|^2 + \alpha_k^2 g_k^2 - \|x_{k+1} - x^*\|^2$$

(Projected) (Sub)gradient Descent

$$\sum_{k=0}^{T-1} 2\alpha_k \langle g_k, x_k - x^* \rangle = \|x_0 - x^*\|^2 - \|x_T - x^*\|^2 + \sum_{k=0}^{T-1} \alpha_k^2 g_k^2$$

$$\leq \|x_0 - x^*\|^2 + \sum_{k=0}^{T-1} \alpha_k^2 g_k^2$$

$$\leq R^2 + G^2 \sum_{k=0}^{T-1} \alpha_k^2$$

All subgradients are bounded in our setting

$$f(\overline{x}) - f^* = f\left(\frac{1}{T}\sum_{k=0}^{T-1} x_k\right) - f^* \leq \frac{1}{T}\left(\sum_{k=0}^{T-1} f(x_k) - f^*\right)$$

Convexity

$$\alpha_k = \alpha^* = \frac{R}{G}\sqrt{\frac{1}{T}}$$

$$\leq \frac{1}{T}\left(\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle\right)$$

Subgradient property

$$\leq GR\frac{1}{\sqrt{T}}$$

$$R^2 = \|x_0 - x^*\|^2, \qquad \|g_k\| \leq G$$

Projected Subgradient Method

$$x_{k+1} = \arg\min_{x \in S} \left(\langle \alpha_k g_k, x \rangle + \frac{1}{2} \|x - x_k\|^2 \right)$$

$$x_{k+1} = \arg\min_{x \in S} \left(\underbrace{f(x_k) + \langle \alpha_k g_k, x - x_k \rangle}_{\text{First order Taylor approximation}} + \underbrace{\frac{1}{2} \|x - x_k\|^2}_{\text{Prox - term}} \right)$$

- The same upper bounds as for the unconditional problem!
- But what if the “local geometry” is not Euclidian?

Mirror Descent

Mirror Descent

$$x_{k+1} = \arg\min_{x \in S} \left(\langle \alpha_k g_k, x \rangle + V_{x_k}(x) \right)$$

$V_{x_k}(x)$ - Bregman divergence (distance) is induced by distance generating function:

$$V_x(y) = \phi(y) - \phi(x) - \langle \nabla \phi(x), y - x \rangle$$

Where DGF is "1" strongly convex w.r.t. primal norm

$$\phi(y) \geq \phi(x) + \langle \nabla \phi(x), y - x \rangle + \frac{1}{2}\|y - x\|^2, \qquad \forall x, y \in S$$

Idea: choose primal norm (with corresponding) dual norm and suitable distance function to fit the geometry of the data

Mirror Descent

$$V_x(x) = 0$$

$$V_x(y) \geq \frac{1}{2}\|x - y\|^2 \geq 0$$

$$\langle -\nabla V_x(y), y - z\rangle = V_x(z) - V_y(z) - V_x(y)$$

TABLE 2.1
Common seed functions and the corresponding divergences.

Function name	$\phi(x)$	$\operatorname{dom}\phi(x)$	$V_x(y)$
Squared norm	$\frac{1}{2}x^2$	$(-\infty, +\infty)$	$\frac{1}{2}(x-y)^2$
Shannon entropy	$x\log x - x$	$[0, +\infty)$	$x\log\frac{x}{y} - x + y$
Bit entropy	$x\log x + (1-x)\log(1-x)$	$[0, 1]$	$x\log\frac{x}{y} + (1-x)\log\frac{1-x}{1-y}$
Burg entropy	$-\log x$	$(0, +\infty)$	$\frac{x}{y} - \log\frac{x}{y} - 1$
Hellinger	$-\sqrt{1-x^2}$	$[-1, 1]$	$(1-xy)(1-y^2)^{-1/2} - (1-x^2)^{1/2}$
ℓ_p quasi-norm	$-x^p \quad (0 < p < 1)$	$[0, +\infty)$	$-x^p + p\,xy^{p-1} - (p-1)\,y^p$
ℓ_p norm	$\lvert x\rvert^p \quad (1 < p < \infty)$	$(-\infty, +\infty)$	$\lvert x\rvert^p - p\,x\operatorname{sgn} y\,\lvert y\rvert^{p-1} + (p-1)\,\lvert y\rvert^p$
Exponential	$\exp x$	$(-\infty, +\infty)$	$\exp x - (x-y+1)\exp y$
Inverse	$1/x$	$(0, +\infty)$	$1/x + x/y^2 - 2/y$

TABLE 2.2
Common exponential families and the corresponding divergences.

Exponential family	$\psi(\theta)$	$\operatorname{dom}\psi$	$\mu(\theta)$	$\phi(x)$	Divergence
Gaussian (σ^2 fixed)	$\frac{1}{2}\sigma^2\theta^2$	$(-\infty, +\infty)$	$\sigma^2\theta$	$\frac{1}{2\sigma^2}x^2$	Euclidean
Poisson	$\exp\theta$	$(-\infty, +\infty)$	$\exp\theta$	$x\log x - x$	Relative entropy
Bernoulli	$\log(1+\exp\theta)$	$(-\infty, +\infty)$	$\frac{\exp\theta}{1+\exp\theta}$	$x\log x + (1-x)\log(1-x)$	Logistic loss
Gamma (α fixed)	$-\alpha\log(-\theta)$	$(-\infty, 0)$	$-\alpha/\theta$	$-\alpha\log x + \alpha\log\alpha - \alpha$	Itakura–Saito

source

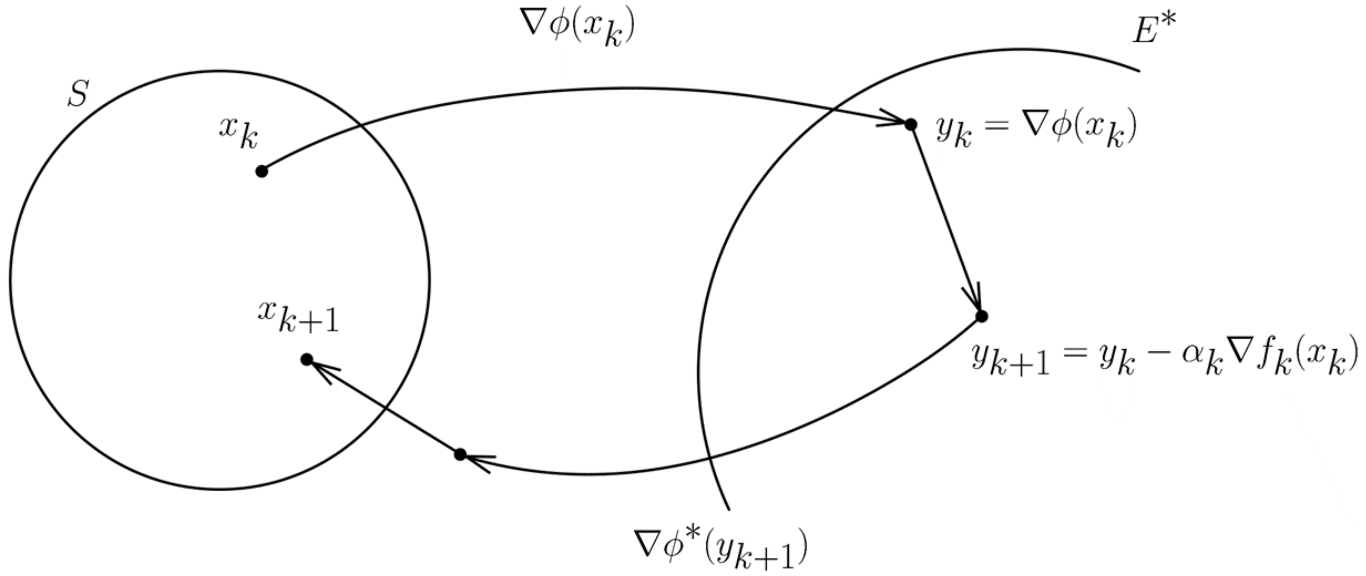
Mirror Descent

$$f(\overline{x}) - f^* \leq \sqrt{\frac{2MG^2}{T}}$$

$$\|g_k\|_* \leq G \qquad\qquad V_{x_0}(x^*) \leq M$$

One more interpretation:

1. $y_k = \nabla\phi(x_k)$
2. $y_{k+1} = y_k - \alpha_k \nabla f_k(x_k)$
3. $x_{k+1} = \arg\min_{x \in S} V_{\nabla\phi^*(y_{k+1})}(x)$



Supremacy

Consider a simple problem, where MD could outperform GD:

$$\min_{x \in S} f(x) \qquad S = \Delta_n = \left\{ x \in \mathbb{R}^n | 1^\top x = 1, x \geq 0 \right\}$$

Choose the primal norm: $\| \cdot \|_1$, corresponding dual norm: $\| \cdot \|_\infty$

$$V_x(y) = \sum_{i \in [n]} y_i \log \frac{y_i}{x_i} = D(y \| x)$$

$$x_0 = (1/n, \dots, 1/n) \;\rightarrow\; V_{x_0}(x) \leq \log n \;\; \forall x \in \Delta_n$$

Supremacy

Let $f(x) = \|Ax - b\|_1$, then $\nabla f(x) = A^\top sign(Ax - b)$

GD

$$f(\overline{x}) - f^* \leq \frac{G_2 R}{\sqrt{T}}$$

$$G_2 = \|A\|_2 \|sign(Ax - b)\|_2 = \|A\|_2 \sqrt{n}$$

$$R = \frac{1}{2}$$

$$f(\overline{x}) - f^* \leq \frac{\|A\|_2 \sqrt{n}}{2\sqrt{T}}$$

MD

$$f(\overline{x}) - f^* \leq \sqrt{\frac{2MG_\infty^2}{T}}$$

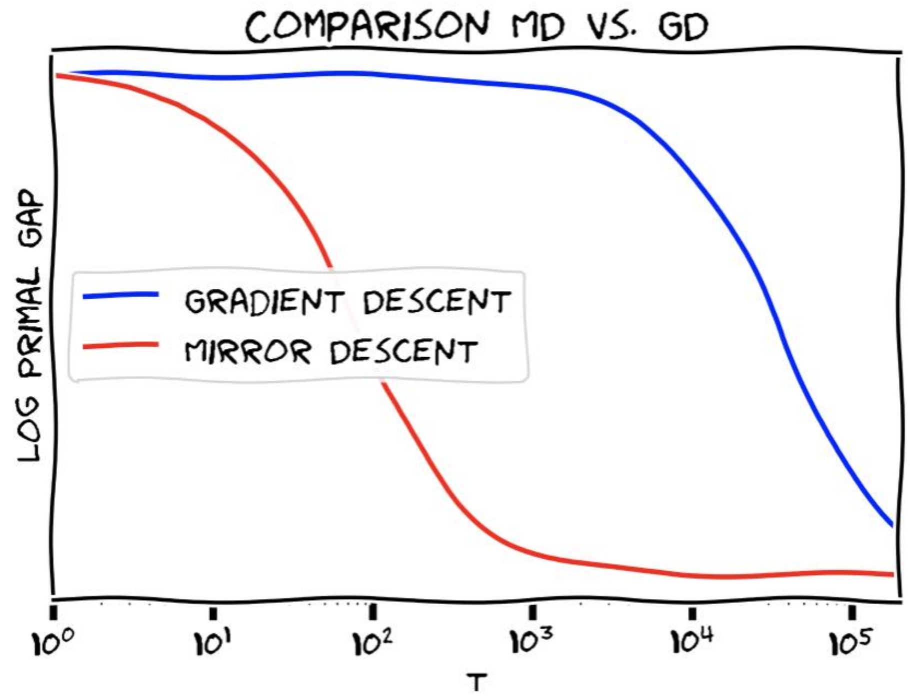
$$G_\infty = \|A\|_\infty \|sign(Ax - b)\|_\infty = \|A\|_\infty \cdot 1$$

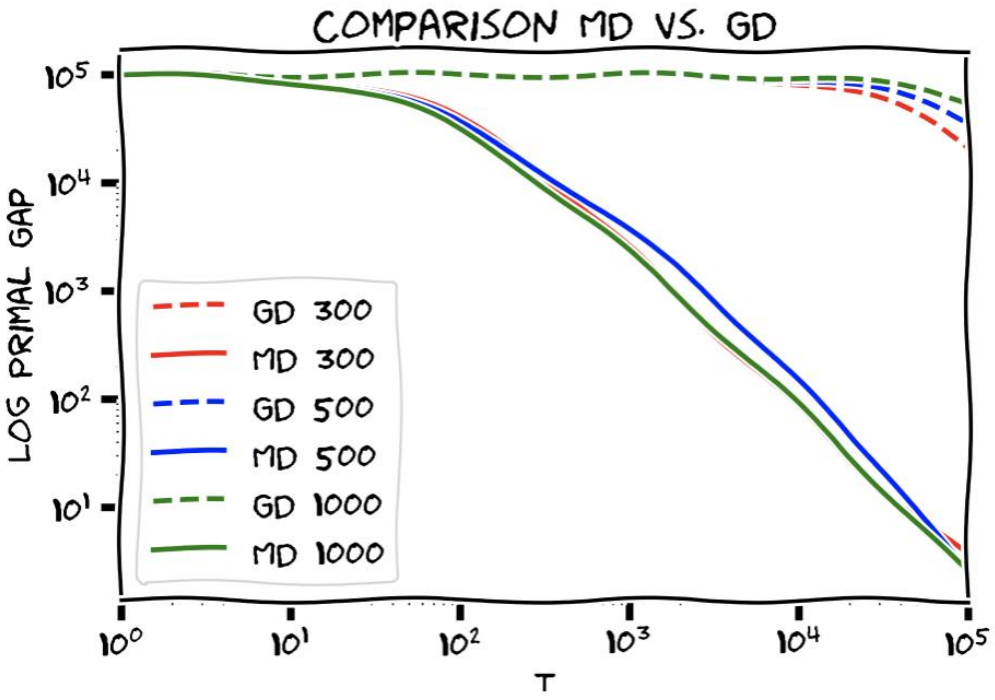
$$M = \log n$$

$$f(\overline{x}) - f^* \leq \sqrt{\frac{2 \log n}{T}} \|A\|_\infty$$

Supremacy

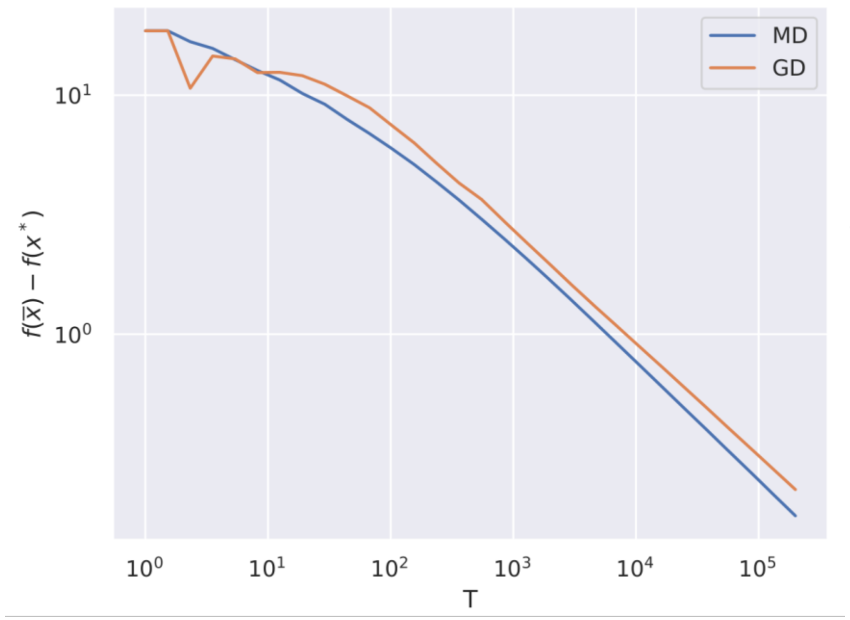
What internet says:



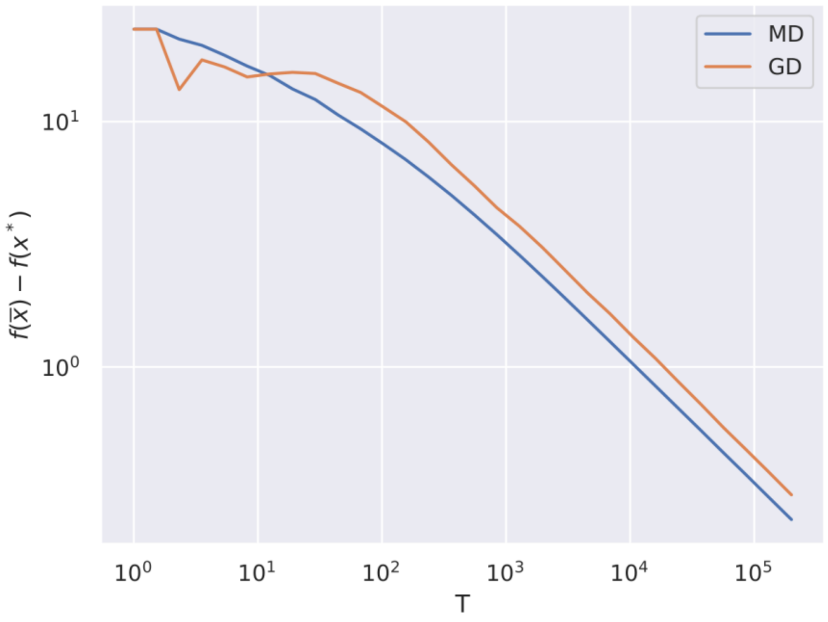


Supremacy

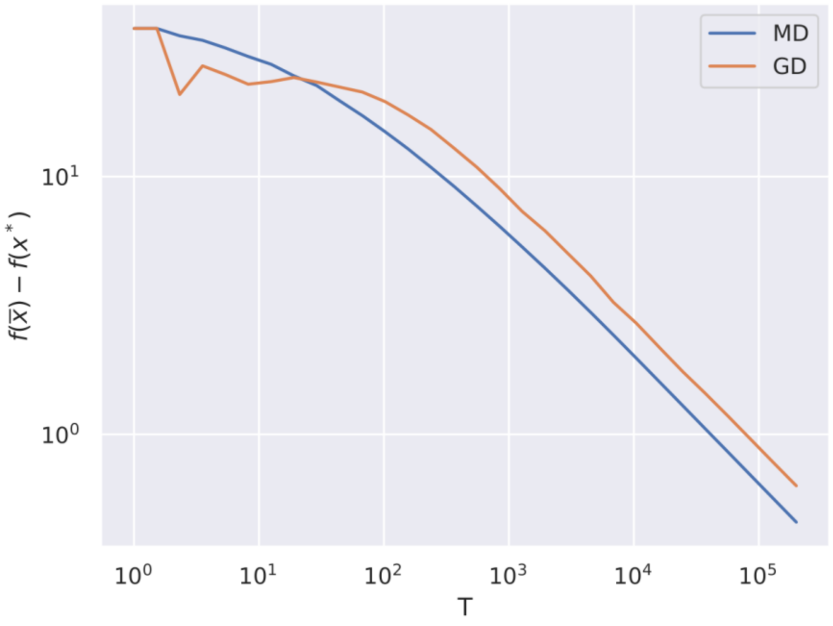
My experiments:



N = 300



N = 500



N = 1000

Around local metric estimation

- Projected subgradient descent $$x_{k+1} = \arg\min_{x\in S}\left(f(x_k) + \langle \alpha_k g_k, x\rangle + \frac{1}{2}\langle I(x-x_k), x-x_k\rangle\right)$$
- (Quasi)Newton methods $$x_{k+1} = \arg\min_{x\in \mathbb{R}^n}\left(f(x_k) + \langle \alpha_k g_k, x\rangle + \frac{1}{2}\langle H_k(x-x_k), x-x_k\rangle\right)$$
- Mirror Descent $$x_{k+1} = \arg\min_{x\in S}\left(f(x_k) + \langle \alpha_k g_k, x-x_k\rangle + V_{x_k}(x)\right)$$
- Natural Gradient $$x_{k+1} = \arg\min_{x\in \mathbb{R}^n}\left(f(x_k) + \langle \alpha_k g_k, x\rangle + \frac{1}{2}\langle (F_k)^{-1}(x-x_k), x-x_k\rangle\right)$$
- Fashionable DL methods:

$$w_{k+1} = w_k - \alpha_k \mathrm{H}_k^{-1}\tilde{\nabla} f(w_k + \gamma_k(w_k - w_{k-1})) + \beta_k \mathrm{H}_k^{-1}\mathrm{H}_{k-1}(w_k - w_{k-1})$$

	SGD	HB	NAG	AdaGrad	RMSProp	Adam
G_k	I	I	I	$\mathrm{G}_{k-1} + \mathrm{D}_k$	$\beta_2 \mathrm{G}_{k-1} + (1-\beta_2)\mathrm{D}_k$	$\frac{\beta_2}{1-\beta_2^k}\mathrm{G}_{k-1} + \frac{(1-\beta_2)}{1-\beta_2^k}\mathrm{D}_k$
α_k	α	α	α	α	α	$\alpha\frac{1-\beta_1}{1-\beta_1^k}$
β_k	0	β	β	0	0	$\frac{\beta_1(1-\beta_1^{k-1})}{1-\beta_1^k}$
γ	0	0	β	0	0	0

Table 1: Parameter settings of algorithms used in deep learning. Here, $\mathrm{D}_k = \mathrm{diag}(g_k \circ g_k)$ and $\mathrm{G}_k := \mathrm{H}_k \circ \mathrm{H}_k$. We omit the additional ϵ added to the adaptive methods, which is only needed to ensure non-singularity of the matrices H_k.

$$\mathrm{H}_k = \mathrm{diag}\left(\left\{\sum_{i=1}^{k} \eta_i g_i \circ g_i\right\}^{1/2}\right)$$

Conclusion

References

References

- Cauchy, A., 1847. Méthode générale pour la résolution des systemes d'équations simultanées. *Comp. Rend. Sci. Paris*, *25*(1847), pp.536-538.
- A. Nemirovksi and D. Yudin. Problem Complexity and Method Efficiency in Optimization. John Wiley Press, 1983.
- Amari, S.I., 1998. Natural gradient works efficiently in learning. *Neural computation*, *10*(2), pp.251-276.
- http://www.pokutta.com/blog/research/2019/02/27/cheatsheet-nonsmooth.html
- http://www.dianacai.com/blog/2018/02/16/natural-gradients-mirror-descent/

References

- http://www.stat.cmu.edu/~ryantibs/convexopt-S15/lectures/24-prox-newton.pdf

SGDR

- https://arxiv.org/abs/1608.03983

Outline

Introduction

Mirror Descent

Conclusion

References

Problems with Adam

- https://arxiv.org/pdf/1705.08292.pdf